# Learning to Recover 3D Human Pose from Silhouettes

**Ankur Agarwal and Bill Triggs**

**LEAR**

**GRAVIR-CNRS-INRIA, Grenoble, France**

**http://www.inrialpes.fr/lear**

# Goal

Recover ***3D human body pose*** from ***image silhouettes***

- 3D pose = joint angles
- use either individual images or video sequences

**Applications**

- motion capture
- human-computer interaction
- action recognition

# “Model Free” Learning Based Approach

- No explicit 3D model — recovers 3D pose (joint angles) by ***direct regression*** against robust silhouette descriptors
- Sparse kernel regressor trained using human motion capture data

**Advantages:**

— no need to build an explicit 3D model

— easily adapted to different people / appearances

— may be more robust than model based approach

**Disadvantages:**

— harder to interpret than explicit model, and may be less accurate

# Prior Work on Pose from Silhouettes

- Brand, ICCV'99: qualitative pose from silhouettes using moment descriptors.
- Mori & Malik, ECCV'02: learn joint centres from broad shape contexts, than use kinematics to recover qualitative pose.
- Shakhnarovich *et al*, CVPR'03: learn quantitative pose using nearest neighbour interpolator (upper body only).

# Image Features

# Why Silhouettes?

- Captures most of the available pose information.
- Relatively simple and low-level, can (perhaps) be extracted automatically from images.
- Assumes no prior labelling of body parts. Insensitive to most surface attributes, clothing colour & texture.

**Limitations**

- ***Artifacts:*** frequently distorted by poor background subtraction / attached shadows, . . .
- ***Ambiguity:*** internal details and depth ordering are hidden

Which arm / leg is forwards?

Where is occluded arm?

Front or back view?

How much is knee bent?

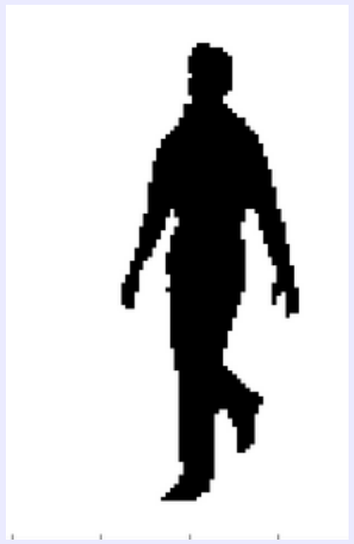
Silhouette-to-pose problem is inherently multi-valued

$\Rightarrow$ Regressors sometimes behave erratically. . .

# Shape Context Histograms

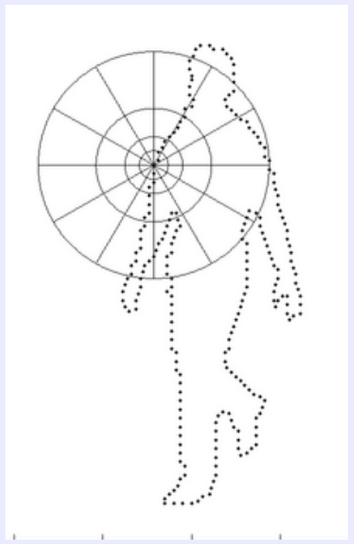
Need to capture silhouette shape but be robust against local occlusions / segmentation failures $\Rightarrow$ avoid global descriptors like moments.

Instead use ***Shape Context Histograms*** — distributions of local shape context responses.

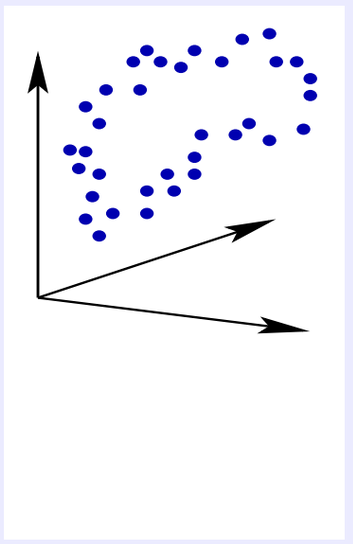


extract silhouette

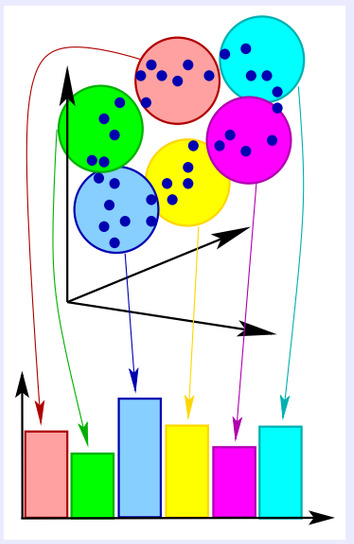
sample edge points



find local shape contexts

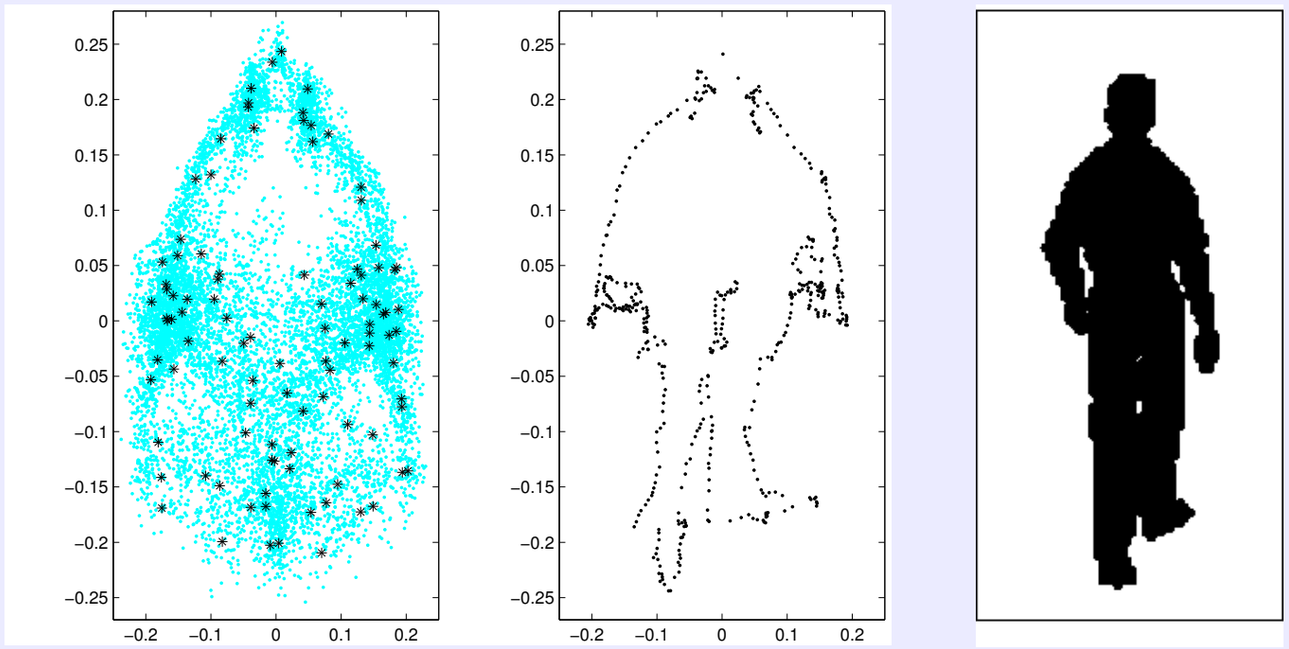


distribution in SC space



vector quantize to get histogram

# Shape Context Histograms Encode Locality




- Left: first 2 principal components of SC Distribution from the combined training data, with $k$-means centres superimposed.
- Centre, right: the SC distribution of a single silhouette.

Appearance is human-silhouette-like as SC's implicitly encode position on silhouette.

# Nonlinear Regression

# Nonlinear Regression Model

Predict system ***output vector*** $\mathbf{y}$ (here a 3D human pose) given system *input vector* $\mathbf{x}$ (here, a shape context histogram):

$$\mathbf{y} = \mathbf{A}\,\mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon}$$

- $\mathbf{f}(\mathbf{x}) = (\phi_1(\mathbf{x}) \;\; \phi_2(\mathbf{x}) \;\cdots\; \phi_p(\mathbf{x}))^\top$ is a vector of scalar ***basis functions*** $\phi_k(\mathbf{x})$
- $\mathbf{A} \equiv (\mathbf{a}_1 \;\; \mathbf{a}_2 \;\cdots\; \mathbf{a}_p)$ is a matrix of learnable ***weight vectors*** $\mathbf{a}_k$
- $\boldsymbol{\epsilon}$ is the residual fitting error.

**Kernel basis**

$\phi_k(-)$ is $k(\mathbf{x}_k, -)$ for given centre points $\mathbf{x}_k$ and kernel $k(\mathbf{x}, \mathbf{y})$.

# Generic Penalized Least Squares

We train the model by adjusting $\mathbf{A}$ to minimize ***squared error*** over ***training pairs*** $\{(\mathbf{x}_i, \mathbf{y}_i) \,|\, i = 1 \ldots n\}$ (silhouettes & 3D poses):

$$\mathbf{A} \;=\; \arg\min_{\mathbf{A}} \left\{ \|\mathbf{A}\,\mathbf{F} - \mathbf{Y}\|^2 + R(\mathbf{A}) \right\}$$

- $\mathbf{x}_i$, $\phi_k(-)$ enter only via ***feature matrix*** $\mathbf{F}_{ki} \;=\; \phi_k(\mathbf{x}_i)$.
- $\mathbf{Y} \equiv (\mathbf{y}_1 \;\; \mathbf{y}_2 \;\; \cdots \;\; \mathbf{y}_n)$
- $R(\mathbf{A})$ is a regularizer / penalty function imposed on $\mathbf{A}$ to control overfitting.

# Damped Least Squares ("Ridge Regression")

For a ***quadratic penalty***, say $R(\mathbf{A}) = \lambda \,\mathrm{trace}(\mathbf{A}^\top \mathbf{A})$, we can solve in closed form using the pseudoinverse (or SVD / QR decomposition / normal equations...):

$$\begin{aligned}\mathbf{A} &:= \arg\min_{\mathbf{A}} \left\| \mathbf{A} \begin{pmatrix} \mathbf{F} & \lambda \mathbf{I} \end{pmatrix} - \begin{pmatrix} \mathbf{Y} & \mathbf{0} \end{pmatrix} \right\|^2 \\ &= \begin{pmatrix} \mathbf{Y} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{F} & \lambda \mathbf{I} \end{pmatrix}^\dagger\end{aligned}$$

# Relevance Vector Machine

- Bayesian-motivated approach to regression & classification.
- Uses a singular ***power-law prior*** to aggressively prune unneeded weights, giving ***sparse solutions*** $\mathbf{A}$.
- RVM typically gives ***very similar performance*** and ***a much sparser solution*** than the corresponding ridge regressor.
- For classification, use the same power-law prior, but minimize logistic error (softmax) not squared residuals.

# The RVM Regularizer

- In regularizer form, the RVM prior is:

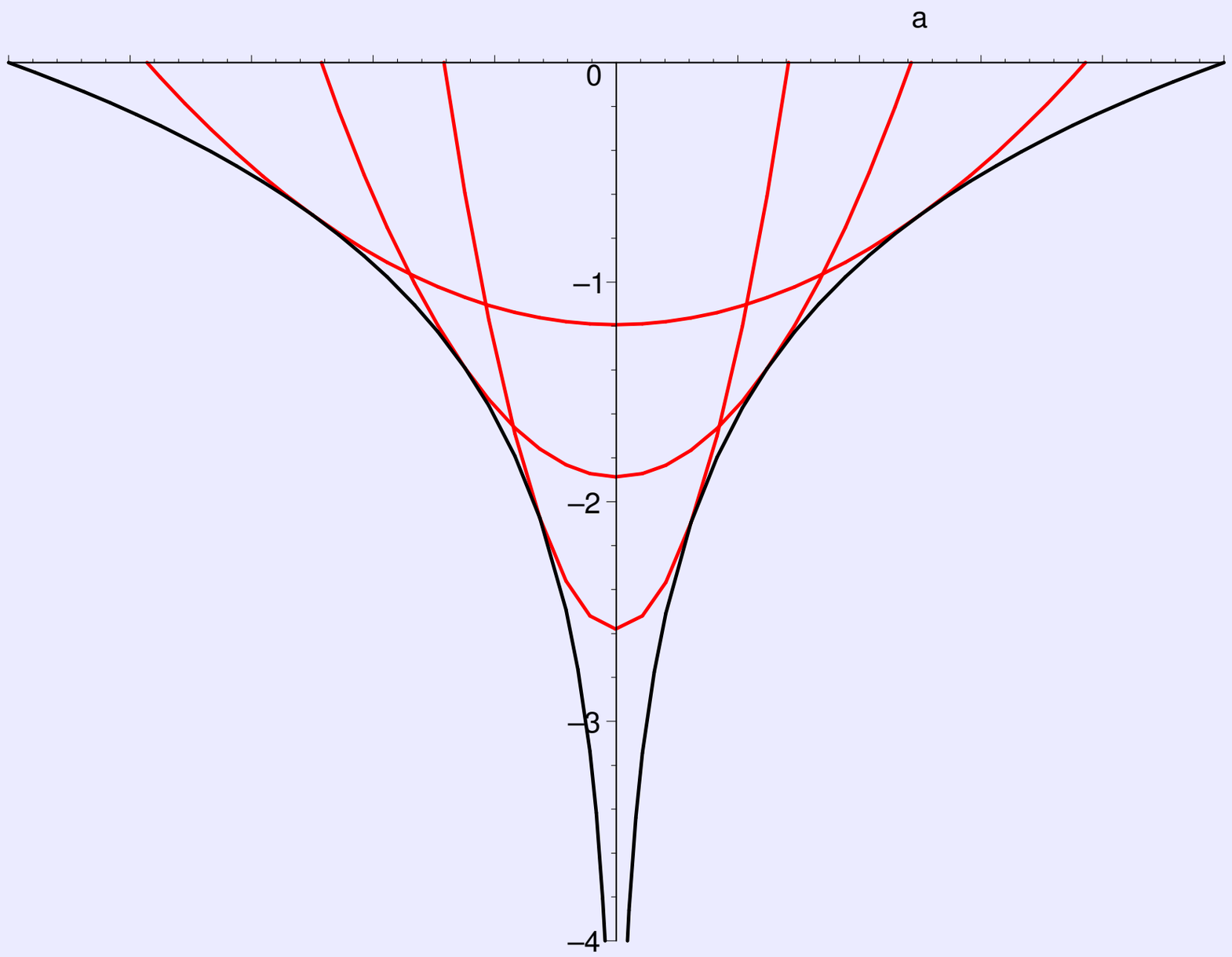
$$R(\mathbf{A}) \;=\; \nu \sum_{a} \log \|a\|$$

- $\nu$ is the pruning / shrinkage strength
- $a$ can be the components, the columns, or the rows of $\mathbf{A}$:
  - columns $\Rightarrow$ prune basis functions $\phi_k$
  - rows $\Rightarrow$ prune (do not fit) inactive output dimensions $\mathbf{y}_i$

# RVM Training Algorithm

The Tipping *et al* hyperparameter-based RVM algorithm doesn't scale well. Instead we use a simple continuation method:

1. Maintain ***running scale estimates*** $a_{\text{scale}}$ for the components or vectors $a$;
2. Approximate the $\nu \log \|a\|$ terms with "quadratic bridges" $\nu \left(a/a_{\text{scale}}\right)^2$ (the gradients match at $a_{\text{scale}}$);
3. Solve the resulting linear least squares problem in $\mathbf{A}$;
4. Update the scales $a_{\text{scale}}$, remove any components that have become zero, and continue until convergence;




The “bridges” prevent premature trapping of components at zero.

# Pose from Static Images

## (CVPR 2004 paper)

# Training & Test Data

- Train on real human motion capture data to capture ***typical*** movements, not just *possible* ones
- For now we don't have the corresponding silhouettes, so we *synthesize* them with Curious Lab's POSER modeller

  — somewhat artificial, but allows a wide range of training viewpoints, plus ground truth for testing.
- Also test on real sequences of another person (no ground truth)

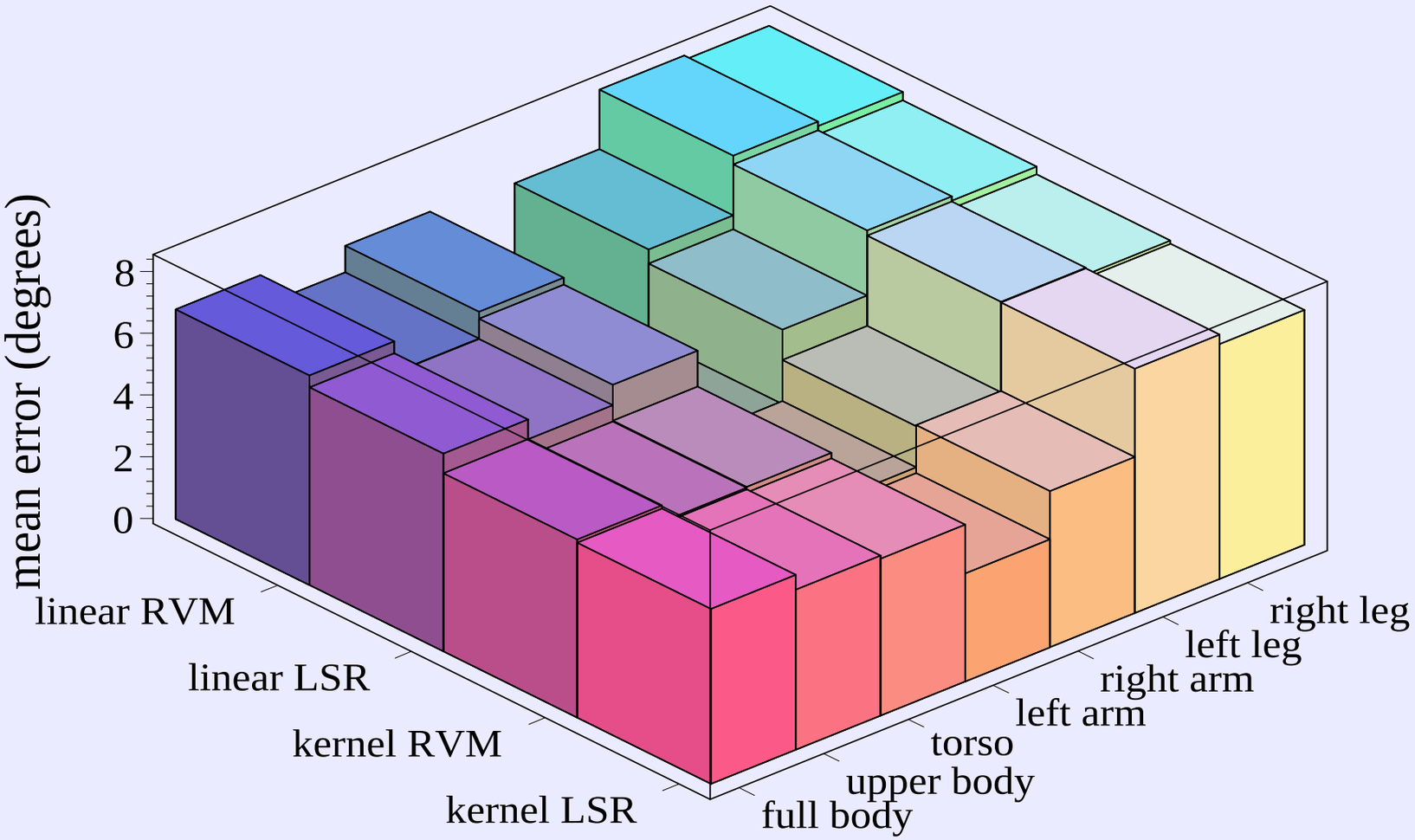
# Methods Tested

***Regressors:*** ridge regression; RVM.

***Basis:*** linear basis (in our nonlinear SC Histogram descriptors); Gaussian kernels of various widths.

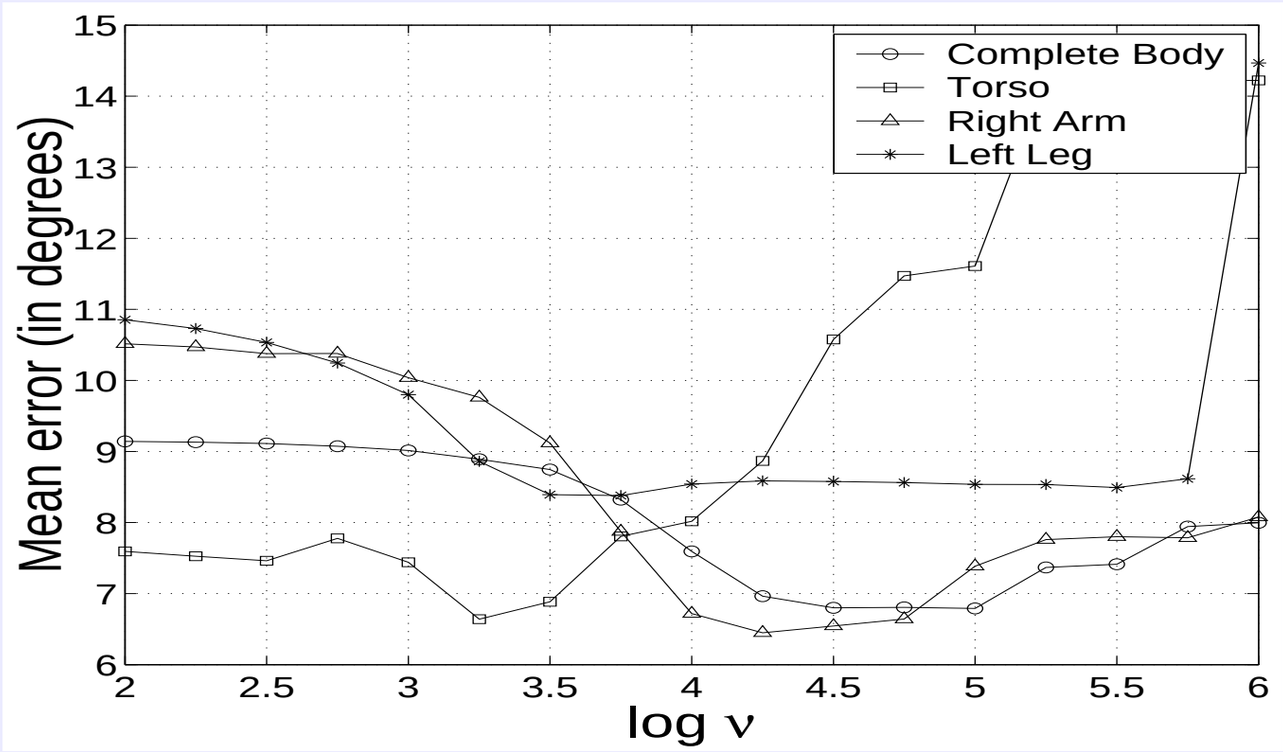
- Performance is very similar for all methods, and insensitive to hyperparameter values.
- Gaussian kernels are only a little better than linear basis.
- The RVM regressors are much sparser than ridge regressors, with very similar performance.

# Regressor Performance




Error versus regressor type and body section for the spiral walking test sequence.

# Residual Error *vs*. Sparsity



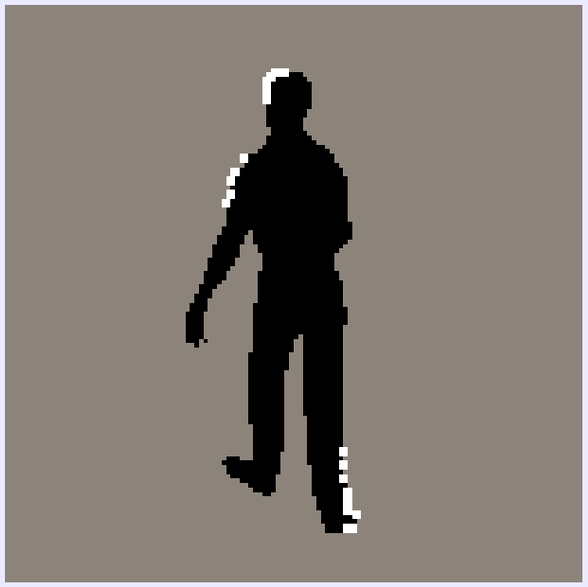

Mean test-set fitting error versus RVM pruning strength $\nu$, for various combinations of body parts.

- Limb regressors are sparser than torso and whole body ones.

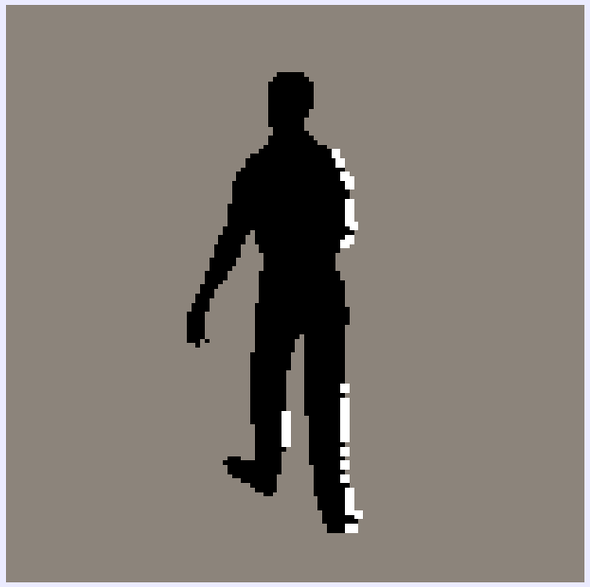
# Relevant Features for Pose Reconstruction

- Owing to SC locality, RVM regressors depend only on isolated regions of the silhouette.
- Training regressors for individual body sections indicates which silhouette regions contribute most to estimating those sections:

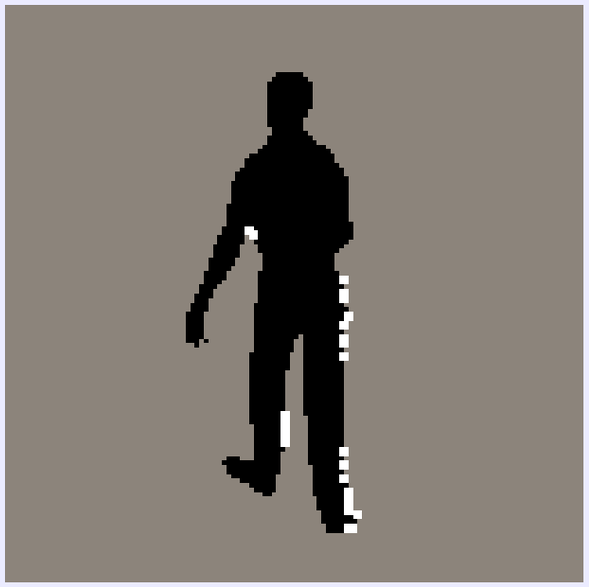
torso & neck



left arm



right leg



Around 10% of the silhouette is used for each section.

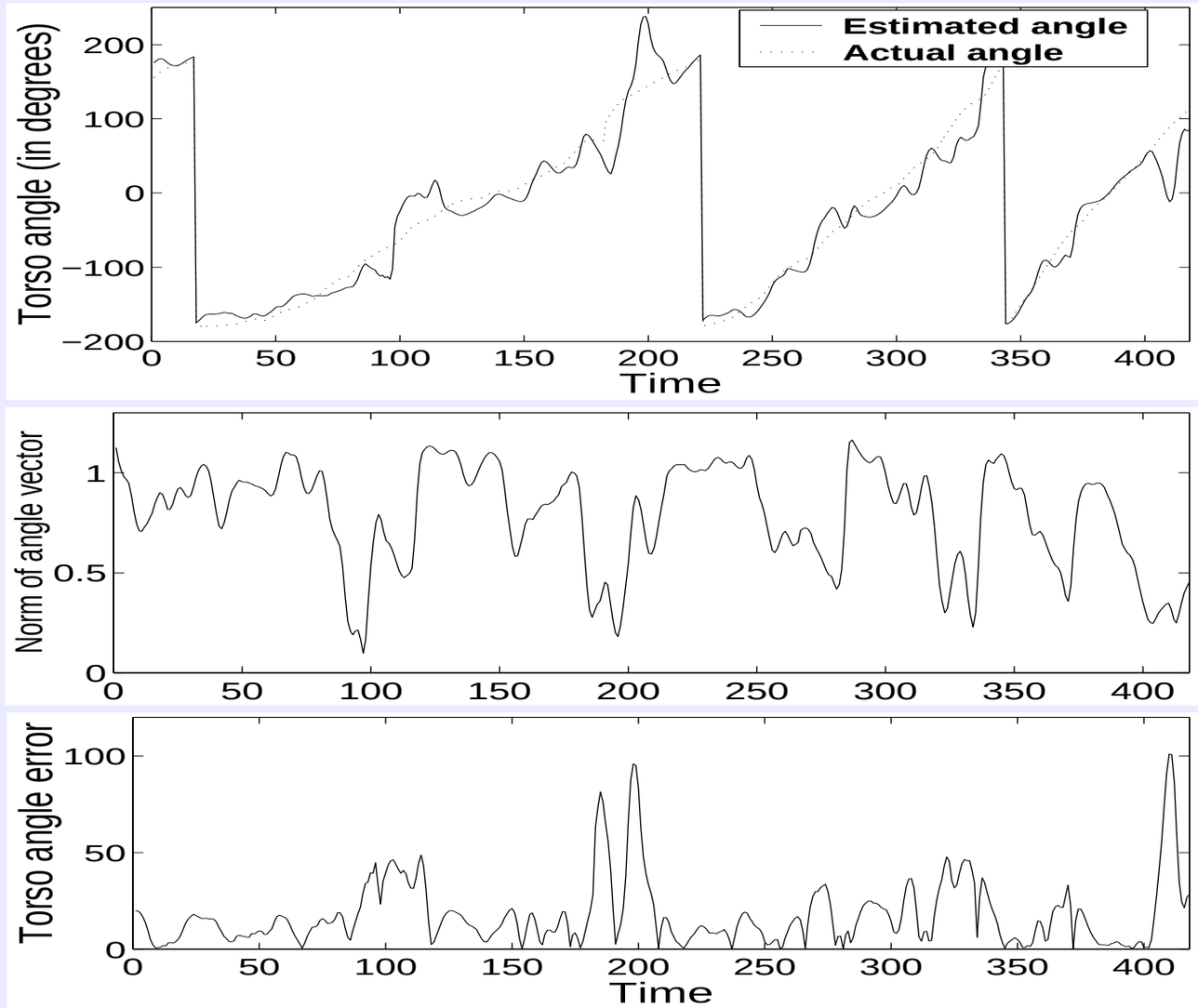
# Synthetic Spiral Walk Test Sequence

- Single image, RVM with Gaussian kernel, sparsity 6% (2636 examples, 156 support vectors).
- Mean angular error per d.o.f. is $6.0^\circ$. *C.f*. (Shakhnarovich, CVPR'03) had $\sim 20^\circ$ and only handled upper body & limited torso rotation.

# Glitches

- Results are OK $\sim 85\%$ of the time, but with frequent "glitches" where regressor either "chooses" wrong case of an ambiguous pair, or remains undecided.
- Especially evident for heading angle — the most 'visible' pose variable.

- For heading, we can quantify the conflict:
  - heading has a $360^\circ$ range so we actually regress $(\cos\theta, \sin\theta)$
  - denormalization of this unit vector is a sign of conflict



# Pose from Video Sequences

(ICML 2004 submission)

# Tracking Framework

- Reduce glitches by exploiting temporal continuity.
- In general, we could use any combination of observations $\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots$ and previous states $\mathbf{x}_{t-1}, \ldots$ as predictors for $\mathbf{x}_t$
- Using observations alone doesn't help much — ambiguities persist for several frames.
- In the end we adopted the familiar filtering framework, (dynamical prediction) + (observation update), but with learned regressive models for both parts
  - — *i.e*. a 'discriminative' approach not generative, model-based one

# State-Sensitive Observation Update

- The observation update model must “know” roughly where the current state lies, so that it can select the correct inverse / corrector mapping to apply

  $\Rightarrow$ The update regressor must depend ***nonlinearly*** on the state prediction

# Regression Based Filtering Equations

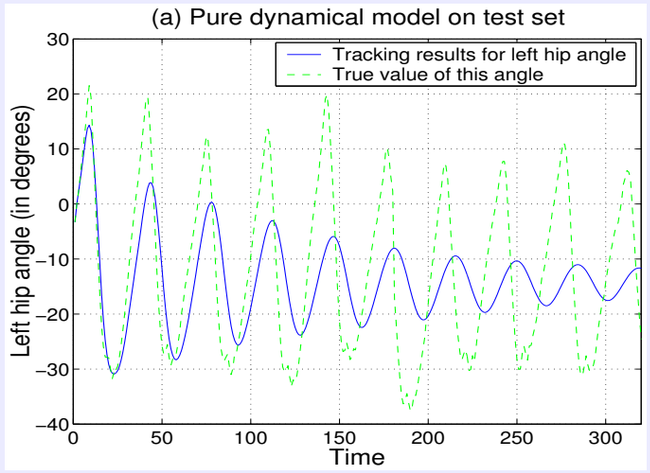
$$\check{\mathbf{x}}_t \equiv (\mathbf{I}+\mathbf{A})(2\mathbf{x}_{t-1}-\mathbf{x}_{t-2})+\mathbf{B}\,\mathbf{x}_{t-1} \quad \text{dynamical prediction}$$

$$\hat{\mathbf{x}}_t \equiv \begin{pmatrix}\mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\check{\mathbf{x}}_t \\ \mathbf{f}(\check{\mathbf{x}}_t,\mathbf{z}_t)\end{pmatrix} \quad \text{observation update}$$

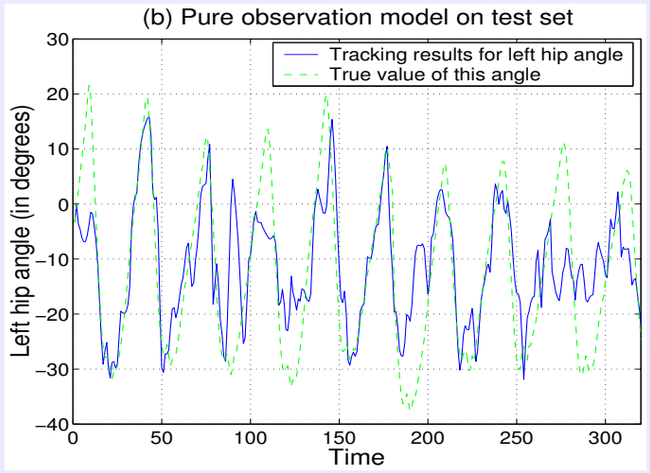
- $f(\mathbf{x},\mathbf{z})$ is a vector of basis functions, *e.g*. kernels.
- $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}$ are weight matrices learned by ridge regression, RVM,. . .

# Dynamics *vs*. Observations — Spiral Walk

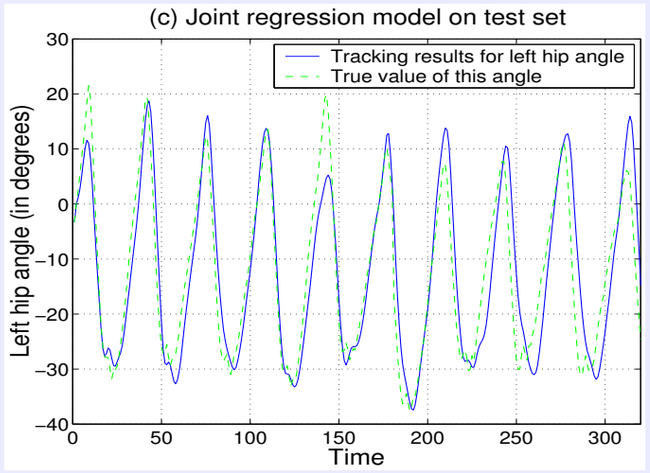
Dynamics only

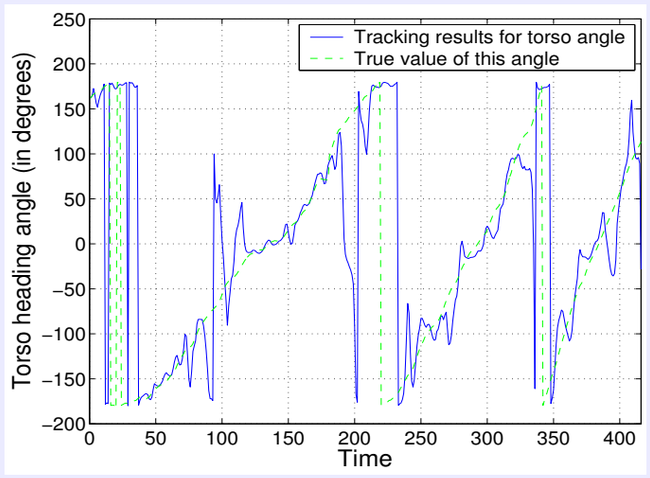


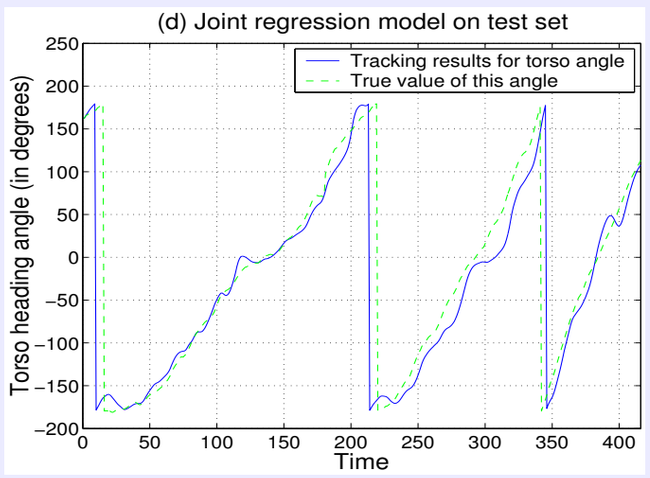

Observation only




Full tracker



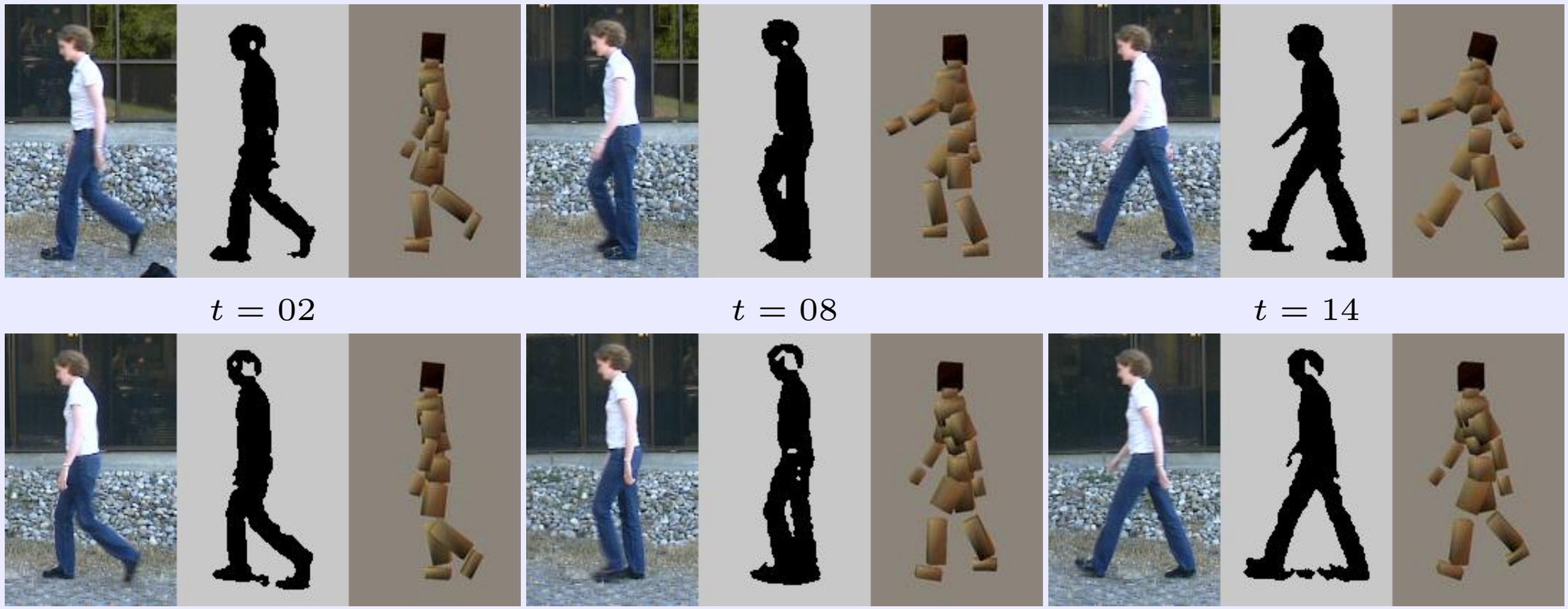
# Examples — Spiral Walk Test Sequence

RVM with Gaussian kernel, 18% nonzero coefficients (1927 training examples, 348 support vectors).

Average RMS estimation error per d.o.f. : $4.1°$.

Resynthesized sequence

# Walking Test, Real Images



$t = 02$ $t = 08$ $t = 14$

$t = 20$ $t = 26$ $t = 32$

# Summary

- "Model free" methods for recovering 3D human pose from monocular silhouettes.
- Direct nonlinear regression of pose against robust histogram of shape context descriptors.
- Ridge regression and Relevance Vector Machine methods, Gaussian kernels.
- Static images and image sequences (regression based filtering).

# The End